Does the Timing of Feedback Given to New Coders of the Communication Complexity Scale Affect Coding Reliability Scores?

By

Danielle Belzer
Bachelor of Science, Speech-Language Pathology, Miami University, 2017

Submitted to the graduate degree program in Speech, Language, Hearing and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

Chairperson: Nancy Brady, Ph.D., CCC-SLP

____

Chairperson: Debby Daniels, Ph.D., CCC-SLP

____

Mindy Bridges, Ph.D., CCC-SLP

____

Date Defended: March 29, 2019

The Thesis Committee for Danielle Belzer certifies that this is the approved version of the following thesis:

Does the Timing of Feedback Given to New Coders of the Communication Complexity Scale Affect Coding Reliability Scores?

Chairperson: Nancy Brady, Ph.D., CCC-SLP

____

Chairperson: Debby Daniels, Ph.D., CCC-SLP

____

Date approved: April 22, 2019

Abstract

The Communication Complexity Scale (CCS) helps quantify communicative behaviors demonstrated by children and adults with disabilities (Brady et al., 2012 & Brady et al., 2018a). When individuals are asked to learn the CCS coding system, they are led through the CCS online training system comprised of instructional modules and independent coding of gold standard videos. Upon completion of scoring, coders receive their scores in comparison to the gold standard scores as well as feedback depicting the reasoning for the score choice. Achieving high reliability is essential for new coders; thus, the feedback provided should be adequate to assist with refining scoring errors. The current study recruited participants to complete the instructional modules and code three gold standard videos according to the CCS protocol. Eight participants were randomly assigned to either the control group or experimental group. The control group completed all tasks as currently arranged by the CCS online training system. In contrast, the experimental group received more frequent feedback on the treatment condition tasks (i.e., two gold standard videos). Then, the experimental group received the same feedback timing as the control group on the follow-up task (i.e., one gold standard video). After the study was completed, the results were calculated via the Mann Whitney *U* test to determine if any statistically significant changes were appreciated. No statistically significant changes between the two groups were noted; although, descriptively, members of the experimental group showed higher reliability on some of the treatment conditions and follow-up tasks. Participants in the experimental group indicated a preference for receiving feedback more frequently to assist with scoring accuracy.

Acknowledgements

First and foremost, I would like to thank Dr. Nancy Brady for her constant support and assistance throughout this process. Without her direction, both the implementation and analysis that this project required would not have been successful. I am fortunate to have had the ability to learn about the intricacies of the Communication Complexity Scale through her guidance. I would also like to thank the other members of my committee, Dr. Debby Daniels and Dr. Mindy Bridges. Through their suggestions and encouragement, we were able to develop an excellent project. Their knowledge of research practices were imperative to the outcome of this study.

Next, I would sincerely like to thank all the students who participated in the project. I am extremely appreciative of the time each of you took out of your busy schedules in graduate school to complete this study. This project would not have been possible without each of you.

I would also like to thank both Kris Matthews and Lisa Hallberg for aiding in the completion of this project. Their willingness to assist with any detail of the project, whether large or small, made the process pan out as smoothly as possible.

Finally, I would like to thank my friends and family for their constant encouragement throughout the completion of this paper and graduate studies. I am privileged to have such a positive support system that continues to be supportive and uplifting.

Table of Contents

Introduction

Communication Complexity Scale

The Communication Complexity Scale (CCS) is a strengths-based measure used to describe the communication skills of individuals with disabilities (Brady et al., 2012; Brady et al., 2018a). Frequently, this scale is used with individuals diagnosed with autism spectrum disorder (ASD), Down syndrome, Rett syndrome, Fragile X syndrome, and Angelman syndrome (Brady et al., 2012). The CCS was developed because there are few measures that quantify the communicative behaviors of those with severe disabilities efficiently and effectively (Brady et al., 2012; Brady et al., 2018a). It is important to create such a measure to not only describe the behaviors of individuals with complex communication needs, but also to better understand their needs for intervention planning (Brady et al., 2012). The CCS scale can help determine whether individuals are progressing along an expected timeline for communication or whether they are showing delays (Brady et al., 2012).

Many individuals with disabilities use a form of communication classified as presymbolic communication or prelinguistic communication. Presymbolic or prelinguistic communication is communication using gestures, body movements, and/or vocalizations (Brady et al., 2012; Brady et al., 2018a). Moreover, presymbolic communication can be further broken down into preintentional communication: "…behaviors that are purposeful but not clearly directed to another person" (Brady et al., 2012) and intentional communication: "…gestures and vocalizations that are clearly directed to another person" (Brady et al., 2012). Following presymbolic communication, individuals begin to use symbolic communication. Symbolic communication is considered the use of real words to express thoughts and ideas (Brady et al., 2012; Brady et al., 2018a).

With the development of the CCS, researchers, clinicians, educators, etc. can now quantify individuals' range of behaviors from presymbolic to symbolic communication (Brady et al., 2012; Brady et al., 2018a). More specifically, the CCS can assist in developing intervention goals and monitoring progress in individuals both for research and practical purposes (Brady et al., 2018a).

The University of Kansas (KU) CCS scripted protocol of activities (henceforth referred to as CCS protocol) is composed of twelve different activities (i.e., wind up, blocks, snack, music, hammer, fan, magna-tiles, dots, bubbles, books, bumble ball, and ball toy) that are intended to give children and adults (i.e., participants) opportunities to communicate their wants, needs, and interests (See Appendix A) (Brady et al., 2012). For example, an activity may include a "sabotaged" broken toy that is given to a child. In contrast, the examiner's toy works functionally, so the child is able to see how the toy operates. The purpose of the activity is to see how the child reacts to the broken toy, potentially by gesturing or vocalizing for assistance (Brady et al., 2012). If the child does gesture or vocalize, this is described as a potentially communicative behavior (PCB).

According to the CCS coding manual, a PCB is described as vocalization or gesture behaviors that are purposeful and communicative in nature (Brady, Matthews, & Muller, 2018b). Vocalizations could include echolalia, consonants and vowels individually or in combination (e.g., CV, CVC sounds), or unintelligible utterances (Brady et al., 2018b). In contrast, gestures completed by the participant could involve both the participant and the examiner. For example, the participant could give an object to the examiner, move the examiners hand towards an object, push the examiner's hand away, tap or touch the examiner, etc. (Brady et al., 2018b). Furthermore, gestures could also be completed by the participant alone. For example, the participant could clap, wave, nod his/her head, shake his/her head, throw objects, tap the table, pantomime gestures, etc. (Brady et al., 2018b). The aforementioned gestures would be considered PCBs if they are completed by drawing attention to an object (e.g., pointing) and/or provide meaning to the interaction by establishing reference to an object (e.g., shaking head) (Brady et al., 2018b).

When coding, behavior regulation and joint attention are defined as functions according to the CCS coding manual (Brady et al., 2018b). These function descriptions are only used for prelinguistic intentional communication and symbolic communication, which translates to scores "6-12" according to the CCS protocol (see below). Behavior regulation is defined as, "A communicative function used to obtain a specific result by either requesting or protesting an object, action, or activity" (Brady et al., 2012; Brady et al., 2018a; Brady et al., 2018b). This could include the participant giving the toy to the examiner to request assistance (e.g., turn the toy on or off, put the toy away, open the container, etc.). It could also include the participant pushing away the toy to reject it, ask for a new toy, protest the activity, etc. (Brady et al., 2018b).

Contrarily, joint attention is defined as, "A communicative function that is used to direct the examiner's attention to an object, event, or activity by commenting, gesturing or vocalizing…" (Brady et al., 2012; Brady et al., 2018a; Brady et al., 2018b). For joint attention, the participant and examiner share an interaction together and engage in social commenting (Brady et al., 2018a). An example of joint attention might include the participant shifting gaze between the object and examiner to show interest in the presented object. During these interactions, the participant may vocalize or gesture to "share" information about the object, but not to ask for more, ask for help, reject it, etc. In these situations, the participant does not change the object, event, or activity to request or protest (Brady et al., 2018a; Brady et al., 2018b).

Coding Communication with the CCS

Scoring. After the entire session is recorded, a coder will later score each individual activity, for a total of twelve scores (see Appendix A for sample score sheet). When scoring, the coder looks for the "highest" function during that specific activity. A score of "0" indicates no response from the child during any point when the object was presented. Scores of "1-5" are considered the preintentional phrase (Brady et al., 2018a; Brady et al., 2018b). Therefore, participant's behavior cannot be inferred as intentional as he/she is only oriented to the examiner or the object (or potentially a small shift in orientation). Scores "1-5" include: single orientation with no PCBs ("1"), single orientation with PCBs ("3-4"), and dual orientation with no PCBs ("5") (Brady et al., 2018b).

Scores of "6-10" are considered to be in the intentional non-symbolic communication phase (Brady et al., 2018a; Brady et al., 2018b). For these scores, there is evidence that the participant is communicating intentionally; however, no words are clearly spoken, signed, or selected on a graphic system. For example, the participant could look between the examiner and the object while handing the examiner the broken toy. It could be inferred that the participant is attempting to request for examiner to fix the broken toy. Scores "6-10" include: triadic orientation with no PCBs ("6") (i.e., participant changes eye gaze from the object to the examiner then back to the object or the participant changes eye gaze from the examiner to the object then back to the examiner), dual orientation with PCBs ("7-8"), and triadic orientation with PCBs ("9-10") (Brady et al., 2018b).

Finally, scores of "11" or "12" are considered to be in the intentional symbolic stage (Brady et al., 2018a; Brady et al., 2018b). During this stage, the participant communicates with one or more spoken words, signs, or augmentative alternative communication (AAC) symbols (e.g., PECS, LAMP Words for Life, etc.). Scores "11-12" include: a one-word sign, symbol, or speech ("11") or multi-word signs, symbols, or speech ("12") (Brady et al., 2018b). See Table 1.1 below for further description.

Table 1.1

CCS Score Description

Number	Definition	Communication level
0	No response	
1	Alerting: a change in behavior, or stops doing a behavior	Pre-intentional
2	Single orientation only on an object, event or person; can be communicated through vision, body orientation, or other means.	Pre-intentional
3	Single orientation only + 1 other PCB	Pre-intentional
4	Single orientation only + more than 1 PCB	Pre-intentional
5	Dual orientation: shift in focus between a person and an object, between a person and an event using vision, body orientation, etc. (without PCB)	Pre-intentional
6	Triadic orientation (e.g. eye gaze or touch from object to person and back)	Intentional Non-Symbolic
7	Dual orientation + 1 PCB (e.g., dual focus + gesture)	Intentional Non-Symbolic
8	Dual orientation + 2 or more PCB (e.g., dual focus + gesture + vocalization, switch closure)	Intentional Non-Symbolic
9	Triadic orientation + 1 PCB (e.g. triadic + vocalization)	Intentional Non-Symbolic
10	Triadic orientation plus more than 1 PCB (e.g. triadic plus vocalization and differential switch closure)	Intentional Non-Symbolic
11	One-word verbalization, sign or AAC symbol selection	Intentional Symbolic
12	Multi-word verbalization, sign or AAC symbol selection	Intentional Symbolic

*PCBs are vocalizations or gestures that are purposeful following the stimulus. They can be completed to communicate behavior regulation or joint attention

Interpreting CCS Scores. As mentioned above, for scores "6-12" (i.e., intentional communication scores), the coder should determine whether the participant's behavior was for behavior regulation (to request) or joint attention (to comment) (Brady et al., 2018a; Brady et al., 2018b). When assessing via the CCS protocol, it is important to determine both the highest levels of functioning and the most consistent levels of functioning for the participants. By defining optimal and typical scores, researchers, educators, clinicians, etc. can observe the communicative abilities the participant has achieved and communicative functions that may develop through proper intervention.

Optimal scores. After the twelve scores are obtained, optimal scores are determined. The optimal score is the average of the top three scores (Brady et al., 2018a; Brady et al., 2018b). If applicable, the average of the top three behavior regulation scores and the average of the top three joint attention scores can also be calculated (Brady et al., 2018a; Brady et al., 2018b). Optimal scores are calculated by finding the three highest scores and dividing that total number by three (Brady et al., 2018b). The purpose of this score is to show the participants' highest level of performance during the activities (Brady et al., 2018a; Brady et al., 2018b). This can also be helpful for further treatment planning when determining preferred items for the participant or communication behaviors the child can achieve (Brady et al., 2018a; Brady et al., 2018b). Additionally, when analyzing the optimal joint attention or optimal behavior regulation scores, individuals can look further into why participants are communicating (Brady et al., 2018b). This, in turn, reveals the participants' strengths while also assisting with intervention planning (Brady et al., 2018a; Brady et al., 2018b).

Typical scores. Similarly, the typical score is calculated after the scoring is complete. In contrast to the optimal score, the typical score is calculated by taking away the top three highest and lowest scores then finding the average of the remaining six scores (i.e., divide the total score by six) (Brady, Matthews & Muller, 2018; Brady et al., 2018). By eliminating the potential outlier scores, other confounding variables such as preferred tasks, nonpreferred tasks, familiarity with the examiner, etc. will be removed from this score (Brady, Matthews & Muller, 2018). The optimal score is important when looking at the participants current level of communication according to the CCS (Brady et al., 2018a; Brady et al., 2018b) Additionally, this score helps researchers and educators discover the communicative functions that the participants are most characteristically exhibiting (Brady et al., 2018a; Brady et al., 2018b).

Reliability

Researchers who study and code videos according to the CCS guidelines must show suitable reliability between coders. Intercoder reliability is defined as, "the extent to which the different judges tend to assign exactly the same rating to each object" (Tinsley & Weiss, 2000). Therefore, coding via the CCS protocol requires new coders to provide the same scores as previous coders that correspond to the participants' communication level. These reliability scores are calculated for all twelve activities. For example, a gold standard video (see below for explanation) formerly scored the behavior of a child during the ball toy activity as a "7" with a function code of behavior regulation. Then, when another coder scores the video as well, he/she would need to code both a score of "7" and a function score of behavior regulation to receive a reliability score of 100% for that activity.

Intercoder reliability is imperative when evaluating consistency between coders. Due to variability among human perception, it is important to improve intercoder reliability because the quality of research depends on the consistency of coding (McHugh, 2012; Tinsley & Weiss, 2000). In an article about interrater reliability, McHugh (2012) discusses when many individuals collect data, they may interpret items differently. This is directly applicable when coding via the CCS protocol as there is a need to analyze measures to ensure that coders' scores have a high percent of agreement with previously established scores. If coders agree on the participant's highest function, it can help make the results of the study more dependable. Additionally, for treatment planning, it is important to ensure those scoring the CCS understand the communicative attempts of those with complex communication needs. When scoring, intervention targets can be obtained from the scores found via the CCS; however, it is essential that the scores are reliable between coders to guarantee proper treatment planning. Although, perfect reliability is rarely achieved between coders, high reliability reflects both the coders' ability to accurately discern behaviors according to the CCS protocol. This also helps determine that the research protocol is testing what it is supposed to be testing (i.e., valid) (McHugh 2012). If the protocol is determined to be reliable and valid, it will assist in proving the results of the study more precise (Tinsley & Weiss, 2000).

As noted by McHugh (2012), studies that incorporate intercoder reliability typically train the data collectors (e.g., coders) and then measure agreement in some capacity (e.g., comparing coders scores to gold standard videos). When new individuals arrive to code per the CCS protocol, they receive online training about the current CCS coding system. Coders are taken through instructional modules that consist of videos, quizzes, and practice coding trials (see "Methods" below for further explanation). Upon completion of the instructional modules, coders are required to code full-length gold standard videos (i.e., videos that include all 12 activities) and then review their reliability scores as compared to the gold standard scores (see Appendix E for examples). In addition to receiving their scores, coders also obtain feedback that consists of a few sentences describing why the gold standard score was chosen (see Appendix F for examples).

Currently, the CCS online training website only provides feedback after coders have completed scoring full gold standard videos. However, prior research has shown that receiving more intermittent feedback positively impacts performance (Alavosius & Sulzer-Azaroff, 1990; Betchtel, McGee, Huitema & Dickinson, 2015; Herzog & Fahle, 1997). It is hypothesized that by providing feedback more frequently, it will influence how coders perceive future scores, thus improving overall reliability scores.

Prior Research on Feedback

Researchers for many years have shown that receiving feedback is imperative for student learning. For example, Herzog and Fahle (1997) found that feedback (especially trial-by-trial and blocked feedback) improved the speed of learning and overall performance. In the trial-by-trial feedback scenario, students received feedback after each incorrect response was recorded (Herzog & Fahle, 1997). In the blocked feedback scenario, feedback was given after a certain number of stimuli were provided to the students, not according to student responses (Herzog & Fahle, 1997). For these two scenarios, feedback was given either more often (i.e., trial-by-trial) or more specific (i.e., blocked) than during the manipulated, partial, or no feedback scenarios. Therefore, students improved on tasks more rapidly with feedback that was provided regularly and/or relayed whether they were accurate or inaccurate in their responses. In terms of CCS protocol coding, it may be helpful to provide more consistent and specific feedback as coders are beginning to learn the CCS coding system. In turn, this could provide more reliable scores as coders continue to score more gold standard videos.

Additionally, the timing of feedback has also been analyzed in practical, real-world settings. One study looked at the timing of feedback for healthcare providers as it related to patient care within the hospital (Alavosius & Sulzer-Azaroff, 1990). It was noted that those who received more timely feedback (i.e., feedback given to workers after one or two tasks were completed each day, rather than weekly feedback) exhibited a quicker acquisition of routines. This demonstrated more recurrent feedback reduced injuries and misconceptions in the healthcare system for new workers. It is hopeful that participants within the current study will produce more reliable coding scores when provided more frequent feedback, similar to that of Alavosius & Sulzer-Azaroff (1990).

More recently, a pilot study was conducted to determine if feedback presented before a task, after a task, or no feedback at all improved performance on a computerized data entry task (Betchtel et al., 2015). While the results determined that there was no difference in performance or speed of skills on data entry, participants indicated a strong preference for any type of feedback over no feedback at all (Betchtel et al., 2015). When reflecting upon how to provide feedback to new coders via the CCS online training website, this knowledge is imperative. While it has previously been determined that coders receive feedback after they submit scores for full videos, this current project will continue to study how the timing of feedback can improve coders' scores.

Prior Research on Coding Videos of Individuals with ASD and Feedback

Recently, one study was completed that reviewed feedback related to video modeling training and feedback for students implementing training procedures for children with ASD (Giannakakos, Vladescu, Kisamore & Reeve, 2015). The participants in this study were age-matched compared to the current study (i.e., 22-year old graduate students) (Giannakakos et al., 2015). In this study, students completed training tasks to learn how to implement effective therapy with individuals who have ASD (i.e., most-to-least prompting hierarchies, least-to-most prompting hierarchies, and prompting delay) (Giannakakos et al., 2015). The researchers determined the video modeling plus feedback was effective in training the graduate students to implement the procedure with individuals with ASD (Giannakakos et al., 2015). Additionally, they found that generalization of skills from the treatment condition (i.e., most-to-least prompting hierarchy tasks) to non-treated tasks (i.e., least-to-most prompting hierarchy &

prompting delay tasks) required less training time due to the feedback implemented in the treatment condition (Giannakakos et al., 2015). While the researchers did not look directly at whether the timing of feedback affected generalization, it is still important to note that direct feedback following a task improved the participants ability to generalize the material. As it has been shown, direct feedback helps develop acquisition of new material (e.g., coding); therefore, for the following study, it is essential to determine if timing of feedback is a variable in increasing the reliability of CCS coding scores.

Purpose of Current Study

This project looked at how feedback timing affected scoring reliability when learning to score according to the CCS protocol. The research question is as follows: "Does the timing of feedback given to new coders affect scoring reliability?". It is hypothesized that the coders who receive the feedback after each activity will have more reliable scoring than coders who receive feedback at the end of scoring gold standard videos.

Methods

Participants

All of the following procedures were approved by the Institutional Review Board (IRB) of the University of Kansas. Members of the Intercampus Program in Communication Disorders (IPCD) at KU were recruited as participants in this study. Students within the program and/or students enrolled in research credits with KU faculty members were recruited for participation. Participating students involved in research with faculty members received credit hours for their participation. Additionally, students enrolled in a first-year IPCD graduate class (Evaluation of Speech-Language Disorders) were also targeted as participants. Participating students from the course received ten extra credit points added to their final course grades upon completion of the study. Information was initially emailed to faculty members (Mindy Bridges, Ph.D., Director of Reading, Language, and Learning Lab; Debby Daniels, Ph.D., Clinical Professor in the Department of Hearing and Speech, Intercampus Program in Communicative Disorders; Kris Matthews, LSCSW, Senior Project Coordinator at Institute for Life Span Studies). Then, faculty members distributed the information to their respective students.

Students who expressed interest were provided a timeline and consent form. After students consented to participate, they were randomly assigned to either the experimental group or the control group; with the exception of the first two participants who were assigned to the control group because they were learning to code as a requirement of their employment in Dr. Brady's CCS research lab. Hence, their instruction needed to be similar to other researchers in the lab. Participants were also provided a participant identification number to protect their personal information and responses throughout the study. In total, nine students consented to participate; eight students completed the study. One participant (identification number: 1006) that consented to participate ceased contact with the investigator and did not complete any tasks for the study. Therefore, this participant withdrew from the study during the initial stages. Refer to Table 1.2 for corresponding identification numbers and groups.

Table 1.2

Participant identification numbers and corresponding groups

Experimental Group	Control Group
1003	1001
1004	1002
1005	1006 (withdrew from study)
1008	1007
	1009

Procedures Overview

This post-test only randomized group design included eight participants that first completed a survey and then reviewed ten instructional modules that presented information about scoring according to the CCS protocol. After the modules were completed, participants tested their coding reliability against a total of three gold standard videos. The control group coded Video 117, Video 205, and Video 402 via the current feedback mechanism that is in place: scoring the entire 12 activity segment and receiving feedback at the end. In contrast, the experimental group coded Video 117 and Video 205 with feedback after each score, creating 12 feedback opportunities during coding. Then, the experimental group coded Video 402 independently as the follow-up measure. Scores between groups were compared for all three videos.

For this study, the independent variable is the timing of feedback (i.e., feedback given after each activity or feedback at the end of the entire video). The dependent variable is the coders' reliability scores for each video (i.e., Video 117, Video 205, and Video 402). The hypothesis was that the experimental group's scores will show stronger reliability when compared to the control group's scores. This will indicate that the more frequent feedback was more effective than the standard feedback.

Survey

A survey was given to each participant to better understand the participants' background prior to the study and to determine if other variables related to scoring potentially affected the outcome of the current study. Data collected in this survey included: their participant identification number, ages, education levels, genders, their prior experiences working with individuals with ASD, and their prior experiences with coding research. See Appendix B below for a depiction of the survey.

Instructional Modules

Participants were instructed to review 10 modules prior to beginning the independent gold standard scoring. First, participants watched a video and then took a short quiz over the foundational module titled, "Introduction to the CCS". Participants learned about what the CCS protocol assesses and why it is an informative tool to use when classifying behaviors of individuals with intellectual and/or developmental disabilities who communicate with minimal verbal skills.

In the second module, participants watched videos and took a quiz regarding how to score the CCS protocol with scores of “0-5”. These scores all indicated the individual's function during that task was in the preintentional phrase (e.g., individual looks at the object OR examiner then vocalizes and/or gestures). In the third module, participants watched videos and took a quiz over how to score behaviors “6-10” or the intentional symbolic phase (e.g., individual looks at the object AND the examiner then vocalizes and/or gestures). Module 4 covered scores of “11” and “12”. Scores of “11” or “12” are considered to be in the intentional symbolic phase (e.g., individual signs, vocalizes words, or uses AAC to communicate).

Next, module 5 covered start and stop times. In this module, participants learned and took a quiz over when to start and stop the video to correspond with the score they were assigning that behavior (learned in the previous module). Module 6 covered communicative functions (i.e., behavior regulation and joint attention) that are used for scores “6-12” or intentional communication scores. In this module, coders learned and were quizzed over how to determine whether the participant's behavior was for behavior regulation (to request) or joint attention (to comment).

Modules 7 and 8 were used to help new coders score with assistance before they are asked to score full videos. In module 7, coders were taken through guided scoring tasks in which activities were scored step-by-step. Then, module 8 covered autonomous scoring in which coders scored a video independently; however, their scores were not recorded for the study.

Module 9 included a video and quiz that covered the scripted administration tasks that were given during recording sessions for scoring. This module merely provided background knowledge so the coders could become familiar with the tasks. Finally, module 10 is where coders began scoring gold standard videos. Overall, the total amount of time it took to complete the modules was approximately nine hours. See Table 1.3 below for description of modules.

Table 1.3

Module Description

Module	Content	Number of Videos	Total Video Length (minutes)	Quiz	Number of Quiz Questions
1	Introduction to the CCS	1	~9:00	Yes	6
2	Scores "0-5"	7	~18:00	Yes	17
3	Scores "6-10"	5	~13:00	Yes	10
4	Scores "11-12"	2	~15:00	Yes	8
5	Start + stop times	1	~7:30	Yes	6
6	Communicative functions	1	~8:00	Yes	13
7	Guided scoring	1	~35:00	No	N/A
8	Independent scoring	1	~24:00	No	N/A
9	Scripted protocol administration	1	~24:00	Yes	9
10	Gold standard videos	N/A	N/A	N/A	N/A

Gold Standard Videos

During the next section, participants were tasked with coding gold standard videos. These videos were considered gold standard because professionals and experienced coders within the lab agreed that these videos represented the most salient scoring. When coders reached the gold standard portion, they were able to see the full video of the interaction between examiner and participant that the coder was scoring. Each video contained 12 activities (i.e., wind up, blocks, snack, music, hammer, fan, magna-tiles, dots, bubbles, books, bumble ball, and ball toy) in which the facilitator presented the toys or object to the participant with the anticipation of facilitating communication. Underneath the video, there was a scoring chart with the 12 activities listed in the order they occurred in the video. For each activity, a start and end time (e.g., 1:10-1:15), score (e.g., “0-12”) and function (e.g., behavior regulation or joint attention for scores “6-12”) drop-down menus were provided (see Appendix C for example). Additionally, an optional comment box was provided for each activity for coders to include notes or questions about the score to be addressed after completion of the study (see Appendix C).

Selection of Experimental Videos

As mentioned above, the experimental group and control group both coded gold standard Video 117, Video 205, and Video 402. These three videos were chosen because they have been ranked "medium-hard" in terms of difficulty from previous coders, according to the confusion matrix (see Appendix D). In other words, some experienced coders scored very high reliability and some experienced coders scored low reliability when coding these three videos.

The confusion matrix compiled and analyzed the reliability of 468 total activities/codes from lab members who coded gold standard videos prior to the study. As depicted below, “true” scores “2-12” are displayed across the top of the document, while “rater” (i.e., coder) scores “2-

12” are displayed down the left side of the document. Then, the dark green boxes displayed in the middle represent correct scores (i.e., the coder scored the activity with the same score as the gold standard). The light green boxes represent scores that were off by one (e.g., the coder scored the activity a “7” when the gold standard was “6” or the coder scored the activity a “5” when the gold standard was “6”). Next, the yellow boxes represent scores that were off by two (e.g., the coder scored the activity a “8” when the gold standard was “6” or the coder scored the activity a “4” when the gold standard was “6”). The orange boxes represent scores that were off by three (e.g., the coder scored the activity a “9” when the gold standard was “6” or the coder scored the activity a “3” when the gold standard was “6”). Finally, the red boxes represent scores that were off by four or more (e.g., the coder scored the activity a “10” when the gold standard was “6” or the coder scored the activity a “2” when the gold standard was “6”).

According to the confusion matrix document, (see Appendix D) many coders struggle when classifying a behavior as a score of "3" vs. a score of "7". The difference between these two scores is determined by the orientation of the participant to the object or the examiner during the activity. If the participant is only oriented to the object or examiner alone, it would be scored a "3". In contrast, if the participant is oriented to both the object and the examiner, it would be scored a "7". When analyzing the red boxes on the confusion matrix (i.e., scores that are off by a margin of four or more), there were 11 instances when coders coded a score “7” when it should have been coded “3”. This is the largest number represented in any of the red boxes, indicating many coders failed to distinguish that the participant in the video was only oriented to the object and not to the object and examiner. Due to this challenging aspect, the videos included in this study incorporated at least one instance in which a coder might accidentally code the function as a "7" when it should have been coded as a "3". Per the confusion matrix, coders scores (i.e., “7”) were matched to gold standard scores (i.e., "3") to determine videos that included instances of this error.

Feedback Conditions

Control group. For the control group, all four participants were instructed to complete the scoring for Video 117 first and Video 205 second, with the standard protocol of scoring gold standard videos (i.e., receiving feedback after completion of scoring full videos). Participants were allowed to use the coding manual and review the completed modules at their leisure during coding. Coding was completed by watching a segment of the recorded sample (i.e., one activity) while looking for the child's highest behavior (i.e., score) during that segment. Once participants believed they found the highest score for that activity, the start and stop times, score, and function (if applicable) were recorded via the CCS online training website. All twelve activities were scored for each video prior to receiving feedback and reliability information. This created a total of 24 items for comparison; 12 items for Video 117 and 12 items for Video 205. Control group participants coded the videos independently and were encouraged to complete the coding for the full video during one occasion, but it was not required.

Experimental group. For the experimental group, participants also coded Video 117 and Video 205 as treatment condition measures (i.e., feedback following each activity). While coding, the investigator (first author) provided feedback after each activity was scored. The feedback provided was the same feedback given to the control group at the end of the video; however, the experimental group received the feedback repeatedly after each task. Participants coded videos while an investigator provided the feedback via Word Document. One page per activity was used to ensure only the feedback from the activity immediately scored was provided. The investigator did not engage, answer questions, or provide any information other than the feedback sentence(s) via the Word Document. Similar to the control group, experimental group participants watched a segment of the recorded sample (i.e., one activity) while looking for the child's highest behavior (i.e., score) during that segment. Once participants believed they found the highest score for that activity, the start and stop times, score, and function (if applicable) were recorded via the CCS online training website. When the participants stated they had finished scoring each individual activity, the investigator displayed the feedback from that activity and allowed time for the participants to review it. This process continued for all 24 treatment condition items (12 for Video 117 and 12 for Video 205).

Determining reliability scores. After each video was scored, the participant's scores were presented in a chart compared to the official gold standard scores. The optimal score, typical score and reliability numbers were provided to each participant via the chart (see Appendix E for examples). Additionally, a brief statement below the chart provided feedback explaining the correct scores (see Appendix F for examples). The experimental group had already seen this feedback via the Word Document during the treatment condition tasks with the investigator. In contrast, the control group was able to review this feedback for the first time, before scoring the next video.

Once the participants completed scoring the two treatment condition videos, they were given two weeks to complete the follow-up measure independently (i.e., Video 402). The control group members coded Video 402 in the same manner they had been exposed to previously (see above for explanation). The experimental group now coded Video 402 without feedback after each score. During this time, no additional feedback or information about coding were provided to the participants. Follow-up scores from the control group and experimental group were compared to determine if the reliability numbers (i.e., participants scores compared to gold standard scores) were higher for the experimental or control group.

Results

Survey

The following table (Table 2.1) represents the responses provided by each participant in the control group.

Table 2.1

Control Group Survey Results

Participant Identification Number	Gender	Age	Highest Degree Received	Years of Experience Working with Individuals with ASD	Type of Experience	Years of Experience Coding for Research
1001	Female	18-30	Bachelor's Degree	1-3 years	Classroom therapy	1-3 years
1002	Female	18-30	Bachelor's Degree	1-3 years	Individual therapy	1-3 years
1007	Female	18-30	Bachelor's Degree	1-3 years	Classroom therapy	N/A
1009	Male	18-30	Bachelor's Degree	1-3 years	Individual therapy, friends/family	N/A

The following table (Table 2.2) represents the responses provided by each participant in the experimental group.

Table 2.2

Experimental Group Survey Results

Participant Identification Number	Gender	Age	Highest Degree Received	Years of Experience Working with Individuals with ASD	Type of Experience	Years of Experience Coding for Research
1003	Female	18-30	Bachelor's Degree	4-6 years	Individual therapy, Nanny	N/A
1004	Female	18-30	Bachelor's Degree	1-3 years	Family/friends	N/A
1005	Female	18-30	Bachelor's Degree	4-6 years	Classroom therapy, Nanny	1-3 years
1008	Male	18-30	Bachelor's Degree	1-3 years	Classroom therapy, Family/friends	1-3 years

As determined by the survey, both the experimental group and the control group showed similar backgrounds prior to the study in terms of age and degree. Overall, the experimental group presented with increased years of experience working with individuals with ASD (total of 10-18 years) in comparison to the control group (total of 4-12 years). All participants had prior experience with individuals who have ASD in some capacity, whether it was as family/friends or in a professional setting (e.g., employment, clinical). Finally, both groups were matched in their previous experiences coding for research. Each group had two participants who had previously had 1-3 years of experience coding and two participants who had no experience coding prior to the study.

Feedback Conditions

The scores from each video were compared to investigate the effectiveness of the feedback mechanisms provided for each group (i.e., were the scores from the participants who had the more frequent feedback better than those with the traditional feedback?). If the results were determined to be statistically significant, the null hypothesis (H_0) was rejected. See below for descriptions of the null hypothesis and experimental hypothesis.

The following table (Table 2.3) displays scoring reliability data Video 117.

Table 2.3

Reliability Scoring Data for Video 117

Participant	Condition	"Score" Agreement Percentage	"Function" Agreement Percentage	"Score + Function" Agreement Percentage
1001	Control	7/12 (58%)	8/9 (89%)	6/12 (50%)
1002	Control	9/12 (75%)	5/9 (56%)	6/12 (50%)
1007	Control	1/12 (8%)	3/9 (33%)	1/12 (8%)
1009	Control	5/12 (42%)	4/9 (44%)	5/12 (42%)
1003	Experimental	6/12 (50%)	6/9 (67%)	4/12 (33%)
1004	Experimental	7/12 (58%)	6/9 (67%)	6/12 (50%)
1005	Experimental	7/12 (58%)	6/9 (67%)	5/12 (42%)
1008	Experimental	8/12 (67%)	8/9 (89%)	8/12 (67%)

The following table (Table 2.4) displays scoring reliability data Video 205.

Table 2.4

Reliability Scoring Data for Video 205

Participant	Condition	"Score" Agreement Percentage	"Function" Agreement Percentage	"Score + Function" Agreement Percentage
1001	Control	6/12 (50%)	9/12 (75%)	4/12 (33%)
1002	Control	6/12 (50%)	8/12 (67%)	4/12 (33%)
1007	Control	3/12 (25%)	2/12 (17%)	0/12 (0%)
1009	Control	6/12 (50%)	9/12 (75%)	5/12 (42%)
1003	Experimental	5/12 (42%)	9/12 (75%)	4/12 (33%)
1004	Experimental	6/12 (50%)	7/12 (58%)	3/12 (25%)
1005	Experimental	3/12 (25%)	8/12 (67%)	3/12 (25%)
1008	Experimental	7/12 (58%)	10/12 (83%)	7/12 (58%)

The following table (Table 2.5) displays scoring reliability data Video 402.

Table 2.5

Reliability Scoring Data for Video 402

Participant	Condition	"Score" Agreement Percentage	"Function" Agreement Percentage	"Score + Function" Agreement Percentage
1001	Control	3/12 (25%)	2/4 (50%)	2/12 (17%)
1002	Control	3/12 (25%)	2/4 (50%)	2/12 (17%)
1007	Control	1/12 (8%)	2/4 (50%)	1/12 (8%)
1009	Control	5/12 (42%)	2/4 (50%)	4/12 (33%)
1003	Experimental	3/12 (25%)	2/4 (50%)	2/12 (17%)
1004	Experimental	2/12 (17%)	3/4 (75%)	2/12 (17%)
1005	Experimental	2/12 (17%)	1/4 (25%)	2/12 (17%)
1008	Experimental	7/12 (58%)	2/4 (50%)	6/12 (50%)

"Score" scores. The total score represents the number of "score" scores (i.e., 12 activities for participants in both the control and experimental group) coded correctly when compared to the reliability (i.e., gold standard) scores.

"Function" scores. The total function score is the correct number of labeled "function" scores (either behavior regulation or joint attention) when compared to the reliability scores. Only scores of "6" or above receive "function" scores (see Introduction section); therefore, the number of functions changed depending on the scores the child received. The child in Video 117 received nine "function" scores providing nine opportunities to assign function. In Video 205, every activity was scored above a "6", leading to 48 total "function" scores for each group. Finally, in Video 402, the child received four scores above "6", totaling 16 "function" scores overall.

"Score + function" scores. The numbers for "score + function" scores refer to the accuracy by which the participants coded both the "score" and the "function" correctly for each activity when compared to the reliability scores. For these scores to be counted accurate, the coders were required to code both the "score" and "function" correctly, as well as labeling joint attention or behavior regulation (i.e., function) appropriately, if applicable. For example, if the gold standard score for the snack activity was a "score" of "8" and a "function" score of joint attention, the participant would only receive perfect reliability if he/she matched by coding a "score" of "8" and a "function" score of joint attention.

Mann Whitney *U* Test. To determine if the differences between the two groups were statistically significant, the differences in the ranks of the outcome measure of the experimental groups and the control group were analyzed. Furthermore, the *U* values for both the control and experimental groups represent the results from the Mann Whitney *U* test (see Table 2.6 below).

When completing the Mann Whitney *U* test for this study, the null hypothesis (H_0) was: There were no statistically significant differences between the reliability scores of the control group and the experimental group based on the timing of feedback given. In contrast, the experimental hypothesis (H_1) was: There were statistically significant differences between the reliability scores for the experimental and control group due to the increased timing of feedback given to the experimental group.

When the sample was assessed, a total of eight scores (four from the experimental group and four from the control group) were assigned ranks. To complete the Mann Whitney *U* test, the total "score" scores, "function" scores, and "score + function" scores were ranked individually on a scale of one to eight from smallest to largest. To explain the intricacies of the Mann Whitney *U* test, the following simplistic example will be used. If Group A receives scores: 10, 11, 12, and 13 and Group B receives scores: 14, 15, 16, and 17, then the rankings would be: one, two, three, and four for Group A and five, six, seven, and eight for Group B. Next, the sums for each group would separately calculated, yielding a rank sum. Following the previous example, the rank sum for Group A would be 10 (i.e., one + two + three + four = 10) and the rank sum from the Group B would be 26 (i.e., five + six + seven + eight = 26). The rank sums would be inserted into the following formula to determine the *U* values: $U_{1\ or\ 2}$ = rank sum - $\frac{M(M+1)}{2}$. The *M* value was the number of data points in each group (i.e., 4 following the previous example). So, the U_1 value for Group A would be 0 (i.e., 10 - $\frac{4(4+1)}{2}$ = 0) and the U_2 value for Group B would be 16 (i.e., 26 - $\frac{4(4+1)}{2}$ = 16).

The following table represents the *U* values calculated via the Mann Whitney *U* Test.

Table 2.6

Mann Whitney U Calculations

Video	Group	"Score"	"Function"	"Scores + Functions"
117	Control	6	3.5	6.5
117	Experimental	10	12.5	9.5
205	Control	8	6.5	8
205	Experimental	8	9.5	8
402	Control	8	8	6
402	Experimental	8	8	10

***p* Values.** The rankings obtained from the Mann Whitney *U* Test were matched to corresponding *p* values using a critical values table for a one-taled test to determine if the results were statistically significant (see Appendix G).

The following *p* values (Table 2.6) were obtained.

Table 2.6

p Values

Video Number	"Score"	"Function"	"Score + Function"
117	.7571	.1000	.2429
205	.3429	.3429	.5571
402	.5571	.1714	.3429

Results from the *p* value analysis yielded no statistically significant differences between the reliability of the control group and the reliability of the experimental group. Thus, we failed to reject the null hypothesis (H_0). The following section will discuss a descriptive analysis of the findings from the study.

Descriptive Explanation of Findings

Although no statistically significant differences were found, most cases reflected increased reliability scores for the experimental group when compared to the control group in terms of “score”, “function”, and “score + function” scores (see Table 2.3, Table 2.4, & Table 2.5).

Video 117. Participants completed Video 117 immediately following completion of the instructional modules. The experimental group received written feedback after each activity was scored, providing 12 feedback opportunities. In contrast, the control group received the same written feedback after all 12 activities were scored (see “Feedback Conditions” under “Methods” above for further description).

For Video 117 (refer to Table 2.3), the highest reliability for the “score” variable was Participant 1002 in the control group with nine of 12 “score” scores from the video recorded accurately. A member of the experimental group (i.e., 1008) was immediately following with eight of 12 correct “score” scores. For “function”, both participants in the experimental group (i.e., Participant 1008) and control group (i.e., Participant 1001) tied for highest reliability by scoring eight of nine “function” scores (i.e., behavior regulation or joint attention) correctly. Finally, in terms of “score + function” scores, a member of the experimental group (i.e., Participant 1008) recorded the highest reliability of all participants with eight of 12 “score + function” scores coded accurately. The overall scores for both “score + function” were notably lower than “scores” and “functions” variables for most participants; however, Participant 1007 in the control group received the lowest reliability for all three variables as noted above.

Video 205. Scoring for Video 205 (refer to Table 2.4) was completed immediately after scoring Video 117. Again, participants in the experimental group received written feedback after each activity, while the control group received written feedback upon completion of scoring the entire video.

For this video, Participant 1008 in the experimental group received the highest reliability scores for each variable (i.e., “score”, “function”, & “score + function”). For “score” scores, this participant accurately coded seven of 12 scores; the next highest reliability was six of 12 “score” scores recorded by both members in the experimental and control group. The scores for “function” were similar, with Participant 1008 scoring 10 of 12 “function” scores correctly; then, participants in both the experimental and control group following with scores of nine. Lastly, for both “score + function” Participant 1008 achieved a score of seven out of 12, recording the highest reliability score by two points when compared to other members of the control and experimental groups. Overall, for this video, most scores for “score” and “function” were comparable to Video 117. The scores for “score + function” were moderately lower for both groups. Similar to Video 117, Participant 1007 received the lowest reliability for all three variables, including zero of 12 accuracy for “score + function”.

Video 402. The final follow-up measure completed by the participants involved coding Video 402 (refer to Table 2.5) independently for both groups. This time, the experimental group did not receive written feedback after each activity; instead, they received feedback after scoring all 12 activities, just as the control group had during the treatment condition videos.

Overall, the participants collectively obtained the lowest scores on Video 402 when compared to Video 117 and Video 205. For “score” scores, again, Participant 1008 achieved the highest reliability with seven of 12 “score” scores. For “function” scores, Participant 1004 in the experimental group received the highest score with three of four “function” scores correctly coded. All other participants with the exception of Participant 1005 in the experimental group received a reliability score of two out of four for “function” scores. Conclusively, for “score + function” scores, Participant 1008 also achieved the highest reliability with six of 12 codes accurately scored. This was followed by Participant 1009 with four of 12. All other participants received scores of two or one correctly coded scores, displaying low reliability overall.

Additional trends. As noted in the above descriptive analysis, most participants’ scores in both the experimental and control group were relatively comparable, removing the highest and lowest outliers. Throughout the study, Participant 1008 in the experimental group coded with the highest reliability when compared to all other participants. In contrast, Participant 1007 in the control group coded with the lowest reliability for all three videos. In many instances, participants’ scores only differed by one or two numbers. For example, in Video 117, the participants in the experimental group received scores of eight, seven, seven, and six. This is similar to many of the other videos and variables as noted in the tables above (Table 2.3, Table 2.4, & Table 2.5).

Moreover, in each video, the “score + function” scores were consistently lower for all eight participants. This is most likely due to the requirement in which the participants must achieve an accurate score for both “score” and “function”. Thus, not obtaining an accurate score for either “score” or “function” impacts the reliability for “score + function”.

Discussion

This study reviewed the feedback conditions provided to new coders of the CCS online training system. Although no statistically significant differences were found via the Mann Whitney *U* test, it is still important to note that members of the experimental group received higher reliability in eight instances when compared to the control group. Per prior research on the frequency of feedback and the results from this study, it is determined that increasing the frequency of feedback given to coders of the CCS online training system may be beneficial in future training for increasing reliability scores.

For the purposes of this study, participants received a relatively short training period prior to coding gold standard videos when compared to typical lab practices. Ordinarily, when new students, assistants, or volunteers arrive, they receive further in-person orientation and training to the CCS protocol. Additionally, participants usually discuss gold standard scores and receive feedback from members of the CCS research lab who have many years of experience coding according to the CCS protocol. In the case of this study, participants coded the treatment condition videos and follow-up video with no other interaction with other study participants and/or CCS research lab members to ensure reliable results. In true practice, coders are able to compare and discuss results, ask questions, etc., especially when first beginning the CCS online training. In addition to these differences, overall, participants from this study did not meet the reliability standards for the CCS research lab. Per conversation with Dr. Brady's, reliability for gold standard videos in the CCS research lab should be greater than or equal to 75%.

Participant Reflections

Upon completion of the follow-up measure, participants from the experimental group reflected on their experience receiving feedback after scoring each activity during the treatment condition (e.g., Video 117 and Video 205) vs. receiving feedback after scoring the full gold standard video (e.g., Video 402). Participants noted that receiving feedback after each activity assisted them in feeling more confident as new coders and helped them determine whether or not they were coding accurately. For example, one participant indicated that it was difficult to determine whether a child's function score was behavior regulation or joint attention; yet, when she received information about why earlier behaviors were qualified as behavior regulation rather than joint attention (or vice versa), she better understood how to score future functions. This is a critical note as receiving feedback on incorrect former scores could alter similar thought processes on latter scores, indicating higher reliability upon completion.

Furthermore, other participants stated that receiving feedback on one score influenced how they coded the subsequent scores. For example, one participant discussed that she had forgotten clapping was considered a PCB. So, after coding a lower score according to the CCS protocol and receiving feedback demonstrating that clapping was a PCB, that impacted coding future activities. Additionally, another participant mentioned she did not see an instance of triadic eye gaze that was noted in Video 117. Yet, when the feedback stated the start and stop times in which the triadic eye gaze was found, the participant was able to watch the video again and better understand how the interaction was scored. This participant also noted that this was particularly helpful because there were more instances of triadic eye gaze scores following the errored one in Video 117.

Survey Responses

On a case-by-case basis, a majority of the participants (i.e., Participants 1001, 1002, 1008) with any years of coding experience (i.e., 1-3 years) showed overall increased coding reliability than participants with no experience coding for Video 117. No differences in coding reliability between participants were noted for Video 205 and Video 402. For example, Participant 1001 had previously had 1-3 years of coding experience prior to the study. Participant 1001 then coded Video 117 with an overall reliability of 63.6%. In contrast, Participant 1007 had a no experience coding prior to the study. This participant demonstrated reliability scores of 15% overall for Video 117. It is possible that those who have been exposed to coding previously may have initially felt more comfortable coding according to the CCS protocol. The discrepancy between participants with prior coding experience and those without prior coding experience was only present during the first video and does not have any significant effects on the outcome of the study. In a future study, this could be controlled by utilizing a regression model (e.g., estimating the relationship among variables) and/or ensuring the participants all have the same (or no) experience coding prior to entering the study.

It was also noted that the individuals with increased experience working with individuals with ASD (i.e., 4-6 years vs. 1-3 years) may relate to an increase in reliability scores. The experimental group had two participants with 4-6 years of experience working with individuals with ASD while the rest of the participants had 1-3 years of experience. It is possible that due to the experimental group having increased experience with individuals with ASD, those participants were better able to apply the CCS protocol when coding videos.

Limitations

The current study was only an initial look into the feedback mechanism used by the CCS online training system. The small sample size (i.e., eight participants) utilized by this study presented a limitation in generalization for further application. Only two individuals who participated in this study continued to code videos as graduate research assistants in the CCS research lab. The majority of the participants had an interest in learning about the CCS protocol and individuals with ASD but will not continue to utilize the scoring mechanism in future practice.

There were also limitations in terms of procedures used throughout the study. Although the control group completed all tasks independently, the experimental group received time in-person with the investigator. While the feedback forms presented to both groups were the same (i.e., written feedback), the experience coding with the investigator observing coding each activity presents a confounding variable of limitation. This could have impacted coding results from members in the experimental group. In the future, to counter this, it would be effective to provide online feedback upon submission of each of the 12 activities.

Implications for Future Research

The current study provides many opportunities for future research and replication. First, replication of this study would be beneficial with an increased number of participants. With more participants, statistically significant conclusions may be drawn to determine if reliability would increase with more frequent feedback.

Additionally, another direction in which concepts from the current study could lead involves the specificities and internal properties of the feedback given. Anecdotally, one participant requested that the feedback include start and stop times for scores within each sentence of feedback, not only in the scoring chart (see Appendix E). This may be a stepping stone to observe the quality of feedback given online, not just the timing and/or frequency of feedback. When observing the quality of feedback, further research could also determine whether online feedback (i.e., the current system) is appropriate for proper learning and increased coding reliability. Researchers could complete another study analyzing in-person feedback training vs.

online feedback training. In turn, this could help decide whether the online system is sufficient for preparing coders to score gold standard videos with high reliability.

In further studies, discussing results, comparing results, and/or asking questions could be reviewed as variables that may impact coding reliability. In the current study, questions were not answered, and results were not discussed with the participants other than the feedback presented upon the completion of activities. In a future study, it may be beneficial to evaluate the quality of feedback (as mentioned above) in terms of answering specific questions and facilitating discussion over results.

From the information gained from the survey, it was shown that experience coding and/or experience working with individuals with ASD may impact reliability scores. In a larger study, variables such as years of experience in the field (e.g., educator or speech-language pathologist) or years of experience coding for research could be reviewed or controlled for to determine if those aspects are correlated to increased reliability outcomes.

Conclusion

Providing feedback is an essential aspect of learning and becoming competent with new skills. Although no statistically significant differences were found in this study, delivering feedback with increased frequency may increase the reliability of individuals coding according to the CCS protocol. Further research is indicated on this subject to confirm this hypothesis. Encouragingly, participants in the experimental group stated positive remarks about the use of feedback after each activity, especially when coding gold standard videos independently. Through the CCS online training website, it may be beneficial to update the current structure to provide feedback after coders score each activity, rather than at the end of each video. In the future, it may be useful to provide in-person trainings that give more specific, direct feedback to new coders. With further research and more definitive findings, the feedback mechanism for coders of the CCS protocol will become more apparent.

References

Alavosius, M.P., and Sulzer-Azaroff, B. (1990). Acquisition and maintenance of health-care routines as a function of feedback density. Journal of Applied Behavior Analysis, 23(2), 151-162. doi:10.1901/jaba.1990.23-151

Brady, N., Fleming, K., Theimann-Bourque, K., Olswang, L., Dowden, P., Sauders, M., and Marquis, J. (2012). Development of the communication complexity scale. *American Journal of Speech-Language Pathology*, *21*, 16-28. doi:10.1044/1058-0360(2011/10-0099)

Brady, N., Fleming, K., Swinburne Romine, R., Holbrook, A., Muller, K., and Kasari, C. (2018). Concurrent validity and reliability for the communication complexity scale. *American Journal of Speech-Language Pathology*, *27*, 237-246. doi:10.1044/2017_AJSLP-17-0106

Brady, N., Matthews, K., & Muller, K. (2018). Communication complexity scale coding manual. Retrieved from https://research.lsi.ku.edu/ccs/manual.php

Giannakakos, A.R., Vladescu, J.C., Kisamore, A.N., and Reeve, S.A. (2015). Using video modeling with voiceover instruction plus feedback to train staff to implement direct teaching procedures. *Behavior Analysis in Practice, 9*(2*)*, 126-134. doi:10.1007/s40617-015-0097-5

Herzog, M.H., and Fahle, M. (1997). The role of feedback in learning a Vernier discrimination task. *Elsevier, 37*(15), 2133-2141. doi:10.1016/S0042-6989(97)00043-6

McHugh, M.L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*. 22(3), 276-282. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23092060

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Tinsley, H., and Weiss, D. (2000). *Intercoder reliability and agreement.* San Diego, CA, US: Academic Press. doi:10.1016/B978-012691360-6/50005-7

Appendix A: Sample Score Sheet

Score Sheet for Scripted Child Assessment: Communication Complexity Scale

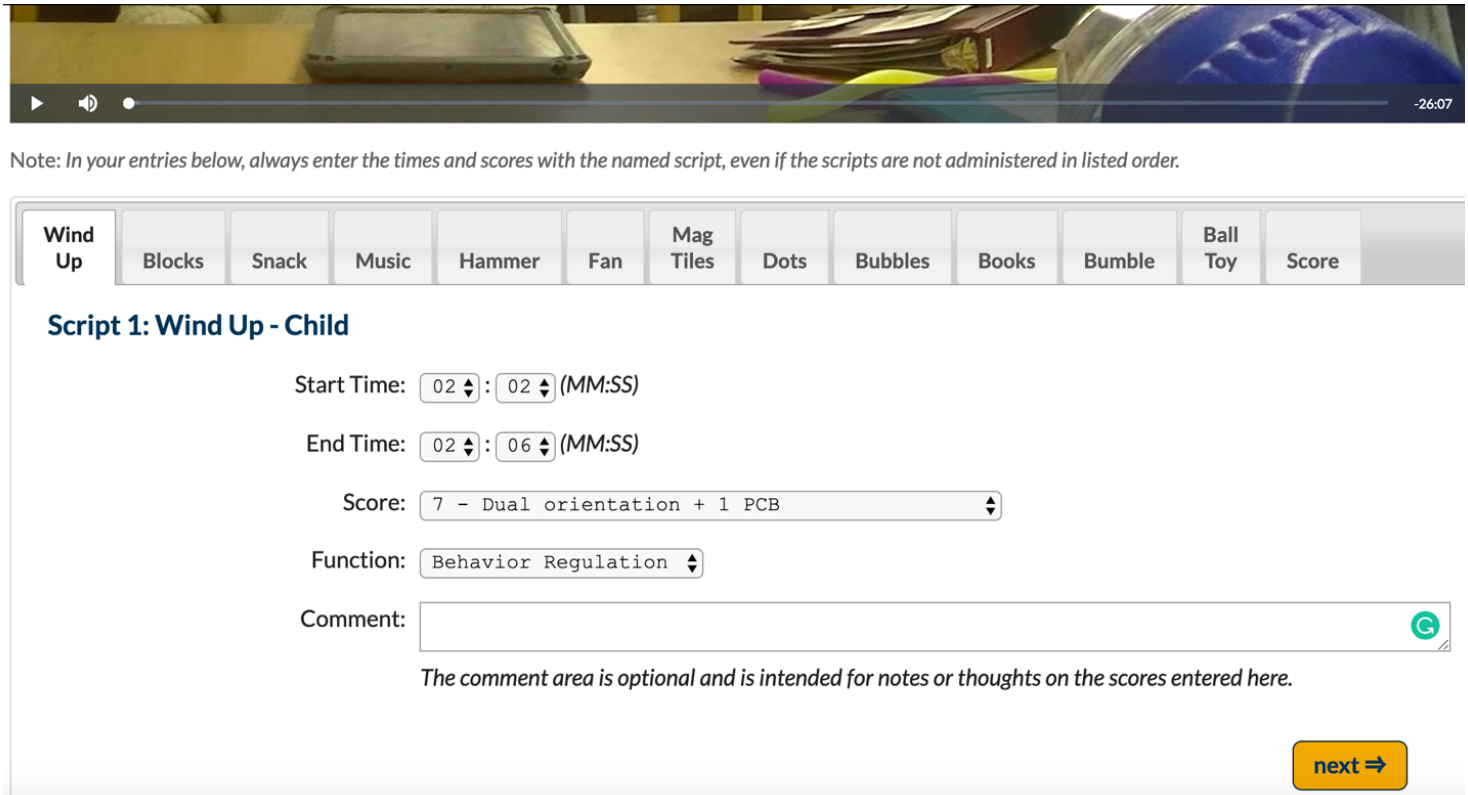
	Participant Behaviors	Wind-Up	Blocks	Snack	Music	Hammer	Fan	M-Tiles	Dots	Bubbles	Books	Bumble	Ball Toy
	Start Time												
	End Time												
	0 No Response												
	1 Alerting												
Pre-Intentional	**2 Single orientation** to person or object)												
	3 Single orientation only + 1 PCB												
	4 Single orientation only + 2 or more PCBs												
	5 Dual orientation dual focus between person and object (without PCB)												
Intentional Non-Symbolic	**6 Triadic orientation** focus from person to object, back to person or object-person-object (without PCB)												
	7 Dual orientation + 1 PCB (e.g. dual focus + vocalization)												
	8 Dual orientation + 2 or more PCB (e.g., dual focus + gesture + vocalization)												
	9 Triadic orientation + 1 PCB (e.g. triadic + vocalization)												
	10 Triadic orientation + 2 or more PCBs (e.g. triadic plus vocalization and differential switch closure)												
Intentional Symbolic	**11 One-word** verbalization, sign or AAC symbol selection												
	12 Multi-word verbalization, sign or AAC symbol selection												
Communicative Function (Scores of 6 or higher)	**Joint Attention (JA)**												
	Behavior Regulation (BR)												
	Response to Question (RQ)												
Communication Modality (Scores of 11 or 12)	**Speech Generating Device (SGD)**												
	Graphic (e.g. PECS)												
	Sign												
	Speech												

Appendix B: Survey

Communication Complexity Scale Feedback Study Survey

1. What is your participant identification number?
   ____
2. What is your gender?
   - Male
   - Female
   - Prefer not to say
3. What is your age?
   - Under 18
   - 18-30
   - 30-45
   - 46-60
   - 60 or older
4. What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.
   - High school graduate, diploma or the equivalent (for example: GED)
   - Some college credit, no degree
   - Associate degree
   - Bachelor's degree
   - Master's degree
   - Professional's degree/Doctorate degree
5. How many years have you been working with people who have autism or are minimally verbal?
   - 1-3 years
   - 4-6 years
   - 7-9 years
   - 10 years or longer
6. Select experiences you have working with people with autism or an intellectual disability
   - ☐ Classroom experience
   - ☐ Individualized therapy
   - ☐ Experiences with family members or friends with ASD/ID
   - ☐ Other ____________________________________
7. Have you ever had experience coding videos for research?
   - Yes
   - No
8. If you responded yes to the previous question, how many years of coding experience do you have?
   - 1-3 years
   - 4-6 years
   - 7-9 years
   - 10 years or longer

Appendix C: Sample Gold Standard Scoring



Appendix D: Confusion Matrix

(updated March 5th, 2018)

ItemScore * GS_ItemScore Crosstabulation

Count

		Gold Standard "True" Score											
		2	3	4	5	6	7	8	9	10	11	12	Total
ItemScore from Rater	2	15	1	0	3	1	0	0	0	0	2	0	22
	3	0	49	1	10	0	2	3	0	0	1	0	66
	4	0	4	1	1	0	3	0	0	0	0	0	9
	5	1	3	1	14	7	5	0	1	1	1	0	34
	6	0	0	0	4	36	7	2	7	1	0	1	58
	7	0	11	1	5	0	57	13	8	2	4	0	101
	8	0	0	2	1	1	5	24	3	3	1	0	40
	9	0	0	0	2	2	8	5	26	6	0	1	50
	10	0	0	0	0	2	4	9	6	7	0	0	28
	11	0	3	1	3	2	4	0	2	0	39	1	55
	12	0	0	0	0	1	0	0	0	1	2	1	5
Total		16	71	7	43	52	95	56	53	21	50	4	468

Appendix E1.1: Scoring Chart for Video 117

	Official scores				Your scores				Agreement		
Script	Score	Func	Mode	Times	Score	Func	Mode	Times	Score	Func	Both
1 Wind_Up	7	BR		01:05 - 01:10	7	BR		01:05 - 01:09	1	1	X
2 Blocks	6	JA		03:25 - 03:30	6	JA		03:22 - 03:28	1	1	X
3 Snack	6	BR		05:00 - 05:05	6	BR		05:00 - 05:05	1	1	X
4 Music	6	JA		07:15 - 07:20	6	JA		07:58 - 08:02	1	1	X
5 Hammer	6	JA		09:29 - 09:34	6	BR		09:30 - 09:34	1	0	
6 Fan	9	JA		10:04 - 10:08	7	RQ		10:36 - 10:41	0	0	
7 Mag_Tiles	2			11:25 - 11:29	2			12:40 - 12:45	1		X
8 Dots	8	BR		15:48 - 15:53	7	BR		15:43 - 15:47	0	1	
9 Bubbles	8	BR		18:38 - 18:43	8	BR		18:37 - 18:42	1	1	X
10 Books	3			21:16 - 21:20	3			21:17 - 21:22	1		X
11 Bumble	8	BR		22:15 - 22:19	8	BR		22:14 - 22:18	1	1	X
12 Ball_Toy	3			24:31 - 24:35	4			24:50 - 24:54	0		
	Avg. Top 3: **8.33**				Avg. Top 3: **7.67**			Reliablility:	**9**	**7/9**	**8**
	Avg. Typical: **6.5**				Avg. Typical: **6.33**				**75%**	**77.8%**	**66.7%**
	Avg. Top BR: **8**				Avg. Top BR: **7.67**						
	Avg. Top JA: **7**				Avg. Top JA: **6**						

Appendix E1.2: Scoring Chart for Video 205

	Official scores				Your scores				Agreement		
Script	Score	Func	Mode	Times	Score	Func	Mode	Times	Score	Func	Both
1 Wind_Up	6	JA		01:39 - 01:45	6	BR		01:42 - 01:46	1	0	
2 Blocks	6	JA		02:40 - 02:44	6	BR		02:55 - 03:00	1	0	
3 Snack	7	BR		04:00 - 04:05	7	BR		04:03 - 04:07	1	1	X
4 Music	7	JA		05:16 - 05:21	6	JA		04:55 - 05:00	0	1	
5 Hammer	7	BR		06:54 - 06:59	7	BR		06:52 - 06:56	1	1	X
6 Fan	6	BR		00:46 - 00:51	6	BR		00:52 - 00:56	1	1	X
7 Mag_Tiles	9	JA		09:02 - 09:07	7	BR		08:27 - 08:31	0	0	
8 Dots	7	JA		10:43 - 10:48	6	JA		10:24 - 10:28	0	1	
9 Bubbles	11	BR	Graphic	12:16 - 12:20	11	BR	Graphic	13:18 - 13:21	1	1	X
10 Books	11	BR	Graphic	14:24 - 14:30	11	BR	Graphic	14:25 - 14:30	1	1	X
11 Bumble	7	BR		16:05 - 16:10	7	BR		16:02 - 16:05	1	1	X
12 Ball_Toy	9	JA		19:19 - 19:24	6	BR		19:30 - 19:34	0	0	
	Avg. Top 3: **10.33**				Avg. Top 3: **9.67**			Reliablility:	**8**	**8/12**	**6**
	Avg. Typical: **7.33**				Avg. Typical: **6.5**				**66.7%**	**66.7%**	**50%**
	Avg. Top BR: **9.67**				Avg. Top BR: **9.67**						
	Avg. Top JA: **8.33**				Avg. Top JA: **6**						

Appendix E1.3: Scoring Chart for Video 402

	Participant Behaviors	Wind-Up	Blocks	Snack	Music	Hammer	Fan	M-Tiles	Dots	Bubbles	Books	Bumble	Ball Toy
#S002	**Start Time**	:56	2:36	5:05	7:35	11:11	12:25	14:15	18:51	22:01	23:51	25:27	30:43
	End Time	1:00	2:41	5:10	7:40	11:16	12:29	14:20	18:57	22:05	23:55	25:32	30:47
	0 No Response												
	1 Alerting - a change in behavior, or stops doing a behavior												
Pre-Intentional	**2** Single orientation only -- on an object, event or person; can be communicated through vision, body orientation, or other means.												
	3 Single orientation only + 1 other PCB (potentially communicative behavior)	**3** voc	**3** voc					**3** voc		**3** voc	**3** voc	**3** PECS	
	4 Single orientation only + more than 1 PCB												
	5 Dual orientation - shift in focus between a person and an object, between a person and an event using vision, body orientation, etc. (without PCB)				**5**				**5**				
Intentional Non-Symbolic	**6** Triadic orientation (e.g. eye gaze or touch from object to person and back)												
	7 Dual orientation + 1 PCB (e.g., dual focus + gesture)												**7**
	8 Dual orientation + 2 or more PCB (e.g., dual focus + gesture + vocalization, switch closure)												
	9 Triadic orientation + 1 PCB (e.g. triadic + vocalization)												
	10 Triadic orientation plus more than 1 PCB* (e.g. triadic plus vocalization and differential switch closure)												
Intentional Symbolic	**11** One-word verbalization, sign or AAC symbol selection			**11** PECS		**11** PECS	**11** PECS						
	12 Multi-word verbalization, sign or AAC symbol selection												
Communicative Function	**Joint Attention**												
	Behavior Regulation			BR		BR	BR						BR
	Response to Question												

Appendix F1.1: Feedback for Video 117

Scoring explanation

Wind-Up: Participant pushes the wind-up towards the examiner (7) though there appears to be a point towards the end of the push, it is an extension of the "push" or give. She appears to be communicating that her toy isn't working, so the function is BR.

Blocks: Participant completes a triadic eye gaze from blocks, examiner, blocks. The examiner talking does not prompt the participant to look back down at the blocks.

Snack: Participant completes a triadic eye gaze from snack, examiner, snack to communicate she needs help opening the snack (BR).

Music: Participant completes a triadic eye gaze from drum, examiner, drum. She does not appear to be requesting or protesting, so the function is JA; she is enjoying the activity with the examiner.

Hammer: Participant completes a triadic eye gaze from hammer, examiner, hammer. She appears to be commenting on the activation of the hammer, so the function is JA.

Fan: The participant reaches toward the fan and completes a triadic eye gaze from fan, examiner, fan. The participant appears to be commenting on the fan, almost as if saying "Hey look at that!" so the function is JA. Another thing that makes it clear that the function is JA is that the participant reaches while the fan is activated. If the fan was off at the time the participant did the communication act, that could show evidence that the function was BR (she wanted the fan turned on).

M-Tiles: The participant's highest act is a single orientation at 12:42. This is the first time in the activity she orients to the magnatiles for at least 3 seconds. At 11:24, she looks at the examiner. However, in order to receive a score of 2 for single orientation to the examiner, the participant must orient to the object/examiner for at least 3 seconds. The participant did not orient to the examiner for this length of time. It cannot be scored as a 5 because the participant was never oriented to an object before/after she looked at the examiner, so there is no dual orientation.

Dots: The participant shows the examiner the marker and pantomimes unscrewing the lid. The examiner saying "what" does break up the communication acts, but because the participant is still showing her the marker after the examiner says "what", the show (after the what) and pantomime are one act.

Bubbles: The participant vocalizes, shows the examiner the bubbles, and looks at the examiner to communicate she cannot open the bubbles.

Books: The participant pushes the book away.

Bumble: The participant shows the examiner the bumble ball and vocalizes.

Ball Toy: The participant vocalizes while oriented to the activity.

Appendix F1.2: Feedback for Video 205

Scoring explanation

Wind-Up: The participant looks from the examiner, to her toy, and back to the examiner. Although it is hard to tell exactly where he is looking when he looks down, it seems as if he is looking at the examiner's toy. Also, he clearly looks from the examiner, down, and back up to the examiner. Because he does not seem to be requesting a new toy, the function is JA.

Blocks: The participant looks from his block, to the examiner, and back to his block. He seems to be communicating about the activity, so the function is JA.

Snack: The participant gives the examiner the snack container. He seems to be communicating that he wants it open, so the function is BR.

Music: The participant is oriented towards his musical instrument through touch and looks at the examiner and vocalizes. Although the vocalization is hard to hear, it is clear with headphones on.

Hammer: The participant gives his broken hammer to the examiner. He does complete a triadic eye gaze before this, but the examiner said "what", which may have caused him to do the give. Therefore, it would be a 7 and not a 9. The function is BR because he is requesting a new hammer/protesting his current one.

Fan: The participant looks from the fan, to the examiner, and back to the fan. The function is BR because he seems to be asking to turn the fan on.

M-Tiles: The participant looks from the magna tiles, to the examiner, and back to the magna tiles while clapping. Since he seems to be roping the examiner into his enjoyment of the activity, the function is JA.

Dots: The participant is oriented towards the squishy toy through touch and looks at the examiner and vocalizes. At 10:46-10:48, it looks like the participant may have completed a triadic eye gaze. However, it is unclear whether the participant actually looks down at the squishy toy or if he is just fluttering his eyes, so only dual orientation is counted. The function is JA because he seems to be commenting about the squishy toy.

Bubbles: The participant points to the "bubbles" PECS, looks at the examiner, and waits. The function is BR because he seems to be requesting about the bubbles.

Books: The participant pushes the book away and taps the Thomas "book" PECS repeatedly. Earlier in the activity, he touched the book PECS, but there was not enough evidence to code it as an 11. At 14:24, he gives enough evidence to show he is intentionally selecting the PECS (pushes other book away, taps repeatedly).

Bumble: The participant pushes the bumble ball over to the examiner. The function is BR because he is protesting the broken one/asking for a different one.

Ball Toy: The participant looks from the balls, to the examiner, and back to the balls while clapping. The vocalization is not coded because the participant is laughing. The function is JA because he is enjoying the activity with the examiner.

Appendix F1.3: Feedback for Video 402

Wind Ups – 3 – 0:56-1:00
Single orientation on toy + vocalization (PCB) = 3. The participant vocalizes while looking at and holding the wind-up toy. The sound at 0:38-0:42 is not counted as a PCB because it is unclear whether it was produced by the participant or by another student off camera. Additionally, the participant's head shaking is not counted as a PCB because it appears to be a stimming behavior.

Blocks – 3 – 2:36-2:41
Single orientation on toy + vocalization (PCB) = 3. The participant vocalizes while holding the block. A score of 5 is not given at 3:38-3:41 because the participant is looking past the examiner, not directly at her.

Snack – 11, BR, G – 5:05-5:10
Use of "goldfish" PECS = 11. The participant gives the "goldfish" PECS to the examiner. The function is Behavior Regulation because he is requesting more goldfish. The mode is Graphic because of the use of a symbol.

Music – 5 – 7:35-7:40
Dual orientation from examiner to toy = 5. The participant looks from the examiner to the shaker in his hand. The initial look up to the examiner is not counted because the examiner's vocalization prompted him to look at her.

Hammer – 11, BR, G - 11:11-11:16
Use of "hammer" PECS = 11. The participant gives the "hammer" PECS to the examiner. The function is Behavior Regulation because he is indicated that his hammer is not working. The mode is Graphic because of the use of a symbol.

Fan – 11, BR, G – 12:25-12:29
Use of the "fan" PECS = 11. The participant gives the "fan" PECS to the examiner. If you listen carefully he gives the PECS when the fan stops, this is his way of requesting for more therefore the function is a BR.

M-Tiles B – 3 – 14:15-14:20
Single orientation on + vocalization (PCB) = 3. The participant vocalizes while looking at the magna tiles. At the same time the participant is banging on the M-tile, however, this seems to be his way of playing with the toy and/or self-stimulatory behavior; thus, it is not coded as an additional PCB.

Dots – 5 – 18:51-18:57
Dual orientation with examiner and marker = 5. The participant has his hand placed on both the examiner and the marker at the same time.

Bubbles – 3 – 22:01-22:05
Single orientation + vocalization (PCB) = 3. The participant vocalizes while looking at the examiner. His laugh right before the vocalization is not scored because laughing does not count as a PCB. It is difficult to determine if the participant looks from the bubbles to the examiner, so it is best to code conservatively and count this as a single orientation, not a dual. between 22:21 – 22:27, the participant pulls his sleeve up while holding the bubble, but this is not a communication act and cannot be considered as a show.

Books – 3 – 23:51-23:55
Single orientation + vocalization (PCB) = 3. The participant vocalizes while looking at the book. A dual orientation is not counted at 24:51-24:56 and at 25:00- 25:04 because the participant is looking past the examiner, not at her.

Bumble Ball – 3 – 25:27-25:32
Single orientation + touch PECS (PCB) = 3. The participant touches the "bumble ball" PECS symbol and begins to pick it up, but then presses it back down. The participant did not give the symbol to the examiner or look at her to "draw her in" so it is not counted as an 11. A score of 11 was not given to the act between 25:48-25:53 where the participant has his hands over the I see PECS because he doesn't grab her attention and does not peruse to communicate.

Ball Toy – 7, BR – 30:42-30:49
Dual orientation + push (PCB) = 3. The participant places his hand on the examiners and pushes hers towards the ball toy. The function is Behavior Regulation because he wants her to put another ball into the toy.

Appendix G: Mann Whitney *U* Critical Values Chart

Table Critical values of the smallest rank sum for the Wilcoxon-Mann-Whitney test

n_1 = number of elements in the largest sample;

n_2 = number of elements in the smallest sample.

Level of significance α						Level of significance α					
Two-sided		0.20	0.10	0.05	0.01	Two-sided		0.20	0.10	0.05	0.01
One-sided		0.10	0.05	0.025	0.005	One-sided		0.10	0.05	0.025	0.005
n_1	n_2					n_1	n_2				
3	2	3	-	-	-	10	6	38	35	32	27
3	3	7	6	-	-	10	7	49	45	42	37
4	2	3	-	-	-	10	8	60	56	53	47
4	3	7	6	-	-	10	9	73	69	65	58
4	4	13	11	10	-	10	10	87	82	78	71
5	2	4	3	-	-	11	1	1	-	-	-
5	3	8	7	6	-	11	2	6	4	3	-
5	4	14	12	11	-	11	3	13	11	9	6
5	5	20	19	17	15	11	4	21	18	16	12
						11	5	30	27	24	20
6	2	4	3	-	-	11	6	40	37	34	28
6	3	9	8	7	-	11	7	51	47	44	38
6	4	15	13	12	10	11	8	63	59	55	49
6	5	22	20	18	16	11	9	76	72	68	61
6	6	30	28	26	13	11	10	91	86	81	73
						11	11	106	100	96	87
7	2	4	3	-	-						
7	3	10	8	7	-	12	1	1	-	-	-
7	4	16	14	13	10	12	2	7	5	4	-
7	5	23	21	20	16	12	3	14	11	10	7
7	6	32	29	27	24	12	4	22	19	17	13
7	7	41	39	36	32	12	5	32	28	26	21
						12	6	42	38	35	30
8	2	5	4	3	-	12	7	54	49	46	40
8	3	11	9	8	-	12	8	66	62	58	51
8	4	17	15	14	11	12	9	80	75	71	63
8	5	25	23	21	17	12	10	94	89	84	76
8	6	34	31	29	25	12	11	110	104	99	90
8	7	44	41	38	34	12	12	127	120	115	105
8	8	55	51	49	43						
						13	1	-	-	-	-
9	1	1	-	-	-	13	2	7	5	4	-
9	2	5	4	3	-	13	3	15	12	10	7
9	3	11	9	8	6	13	4	23	20	18	14
9	4	19	16	14	11	13	5	33	30	27	22
9	5	27	24	22	18	13	6	44	40	37	31
9	6	36	33	31	26	13	7	56	52	48	44
9	7	46	43	40	35	13	8	69	64	60	53
9	8	58	54	51	45	13	9	83	78	73	64
9	9	70	66	62	56	13	10	98	92	88	79
						13	11	114	108	103	93
10	1	1	-	-	-	13	12	131	125	119	109
10	2	6	4	3	-	13	13	149	142	136	125
10	3	12	10	9	6						
10	4	20	17	15	12						
10	5	28	26	23	19						